

eBooks with Indexes that Reorganize Conceptually

Ed H. Chi, Lichan Hong, Julie Heiser, Stuart K. Card

Palo Alto Research Center

3333 Coyote Hill Road, Palo Alto, CA 94304

{echi, hong, heiser, card}@parc.com

ABSTRACT

Subject indexes were an important step forward for books because they enabled the comparison and correlations of information without extensive reading, re-reading and memorization. In this short paper, we focus on the user interaction and usage scenario of a new system called ScentIndex that enhances the subject index of an eBook by conceptually reorganizing it to suit particular information needs. Users first enter information needs via keywords describing the concepts they are trying to retrieve and comprehend. ScentIndex then computes what index entries are conceptually related, and reorganizes and displays these index entries on a single page.

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical User Interfaces; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia—Navigation; User Issues H.5.m. [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

General Terms: Design, Human Factors.

Author Keywords

Book Index, eBooks, Information Scent, contextualization, personalized information access.

INTRODUCTION

Representing a book electronically on the computer is part of the long effort by researchers seeking to enhance the reading activity [1, 4, 5, 6, 11, 14, 16, 17]. One major advantage of the electronic equivalent of a paper book is its ability to offer searching mechanisms of various kinds to the user. Keyword searches have become a mainstay of many eBook products. However, in the process of creating new electronic indexes, eBook system designers have thus far neglected to take advantage of existing manually generated subject indexes. This is surprising, since whole fields have been devoted to the understanding of how to generate a good subject index for a book manually.

Latest efforts in digital libraries such as the Amazon.com “Search Inside the Book™” [3] and the Million Book Project [9] have focused on the preservation and digitization of the vast archive of paper books that human efforts have accumulated. Many, if not nearly all, of these

books contain subject indexes that have been painstakingly generated manually at the time of the production.

In this short paper, we focus on describing the user interaction model and a usage scenario of our new ScentIndex technique, which conceptually reorganizes a subject index. A companion manuscript describes the actual algorithm and a user study on the effectiveness of the technique [7].

The ScentIndex technique works in the following way. The user first enters a set of keywords that describe the concepts that she is interested in locating in the book. The system then quickly narrows down a large index containing thousands of entries to tens of extremely relevant entries and displays them on a single page for the user to peruse. In our method, we use a word co-occurrence matrix [15] to extract index entries that are conceptually relevant to the keywords the users entered.

We also enhance the navigation and browsing interactions between the index and the eBook. We use a wide variety of highlighting techniques to give navigational cues to the users. These cues tell users: (1) what index entries are likely to satisfy their query, and (2) what passages on the page are likely to contain the relevant information pieces. These simple navigational cues enable users to check out multiple index entries rapidly.

SuperBook [13] is probably the most well-known related work to the ScentIndex, because it provides a table of content (TOC) that is hierarchically and dynamically presented, but not for the subject index. Its algorithm is based on term frequency, whereas ours is based on term conceptual similarities. Also, instead of seeking to automatically generate new indexes for book texts as in [10, 12], we are interested in ways of enhancing these existing indexes so that they reorganize themselves to better suit the information needs of the user.

USAGE SCENARIO AND USER INTERACTION

ScentIndex is designed as a component of an electronic book called 3Book [6]. Figure 1 depicts the process by which the eBook is produced. First the paper book is scanned and OCR’ed, producing page textures used in the 3D graphics texture mapping process. The word locations are extracted to enable highlighting of the individual words. The recognized text is then used to compute the word association matrix. We use the matrix to compute the

Degree-Of-Interest (DOI) function for the ScentIndex, ultimately producing a single page of conceptually relevant index entries. The DOI function is computed based on word co-occurrence and Information Scent [8].

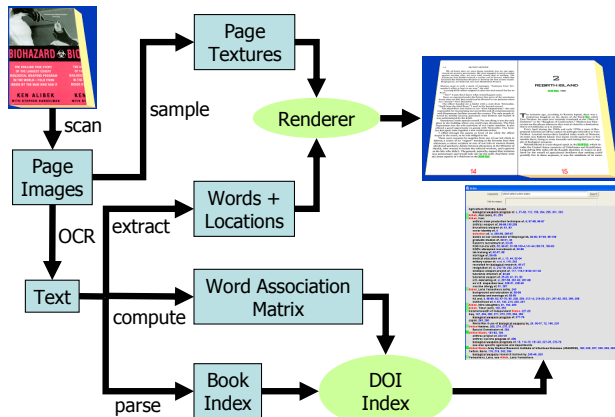


Figure 1: Overall flowchart.

The user study and the scenarios below are based on the book *Biohazard* by Ken Alibek [2], which is a non-fiction retelling of his experiences working on biological weapons in the former Soviet Union. There were 13 index pages in two columns, consisting 829 entries, which would have been impossible to present fully on the screen at one time.

In the following description of the system, we use one of the comparison tasks in our user study to demonstrate the user interaction with the ScentIndex. The task is “*What year did Russia open negotiations with Iraq for large fermentation vessels? What year did Vladimir Kryuchkov become chairman of the KGB? Which occurred first?*”

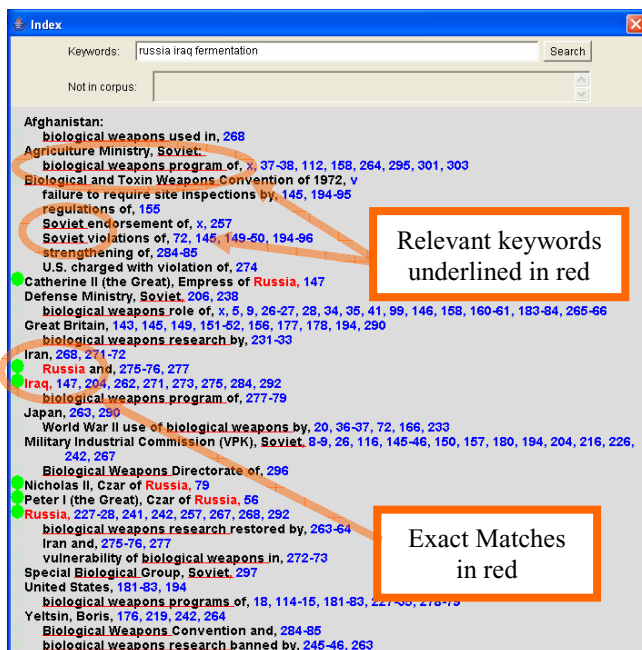


Figure 2: ScentIndex showing the reorganization after "russia iraq fermentation" was entered as the information need.

Figure 2 shows the ScentIndex after the index entries have been reorganized according to the information need of “russia iraq fermentation”. We see keywords highlighted in red showing *exact* keyword matches. Relevant words such as “biological”, “weapons” are highlighted by underlining.

Figure 3 describes how the user interacts with this index view. After the user has specified the concepts in the keyword box, the method computes a new single-page index view. The user can then click on a page number associated with an index entry, which opens the book to that particular page. User query keywords and the words in that index entry are used for keyword highlighting on the book page.

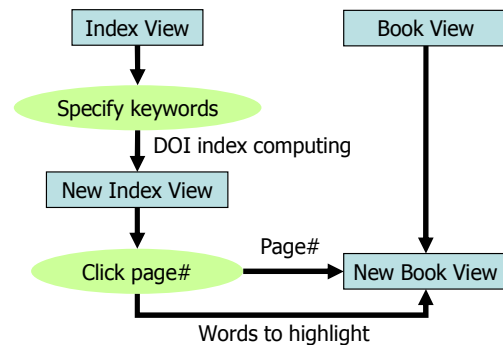


Figure 3: Flow chart describing the user interaction.

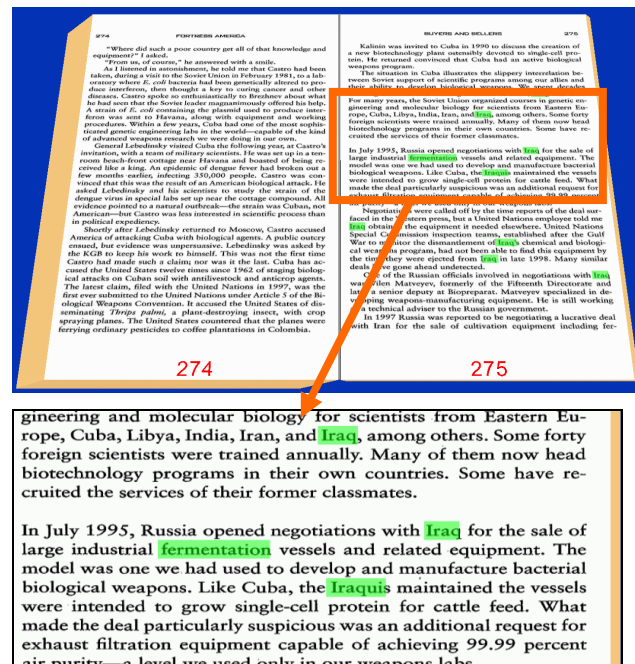


Figure 4: Clicking on the 6th "Iraq" page number 275, the book opens up to that page (top) and highlights the relevant keywords (detailed bottom).

As shown in Figure 2, there are many entries that are relevant to the query. The user decides that the “Iraq” entry is the most relevant because, since the whole book is about biological weapons and Russia, the other entries do not

stand out. The user successively skims the page entries under “Iraq”. Figure 4 shows the book turned to a new page describing the 6th page entry of “Iraq”. The words “*russia iraq fermentation*” were automatically highlighted on this page. The bottom figure shows the specific paragraph that gives the first answer. The key to finding the answer quickly is looking for a page that contains all three of these words in a single passage. As shown, the automatic highlighting enables the user to quickly find the passage that might be relevant to the user query. This highlighting is extremely helpful in tasks that require skimming for related facts through a large list of pages. The answer is 1995.

For the second part of the question, we need to find out what year *Vladimir Kryuchkov became chairman of the KGB*. Figure 5 shows the ScentIndex after the query “kryuchkov chairman kgb” is entered, thus eliminating hundreds of index entries. There are still many potentially relevant entries, but “Kryuchkov, Vladimir” is probably the most relevant. Figure 6 shows the results after clicking on the 2nd page entry. We see the answer is 1988.

After completing these two parts to the question, the user can then compare the two facts and find the 2nd fact occurred first in 1988. Readers should note that there are other ways to navigate through the index to find the solution to this question. We have merely shown one possible way to locate the relevant information.

The tasks here seem easy for a number of reasons: (1) First, by using conceptual reorganization, users have a high confidence that relevant entries are not omitted, because we do not rely solely on exact keyword matches. It is known that users generally have a hard time formulating a good set of search keywords, as there are large subject variations in formulating search queries in search engines, even when the task is explicitly specified. For the first question above for example, without the reorganization and fitting the index on a single page, a user might have first looked under other potential entries such as “Russia” and “fermentation” without success. Instead, users are able to decide “Iraq” is the most promising entry.

(2) Second, there are many page entries that are potentially relevant. Our keyword highlighting on the book pages helped in skimming for relevant passages. By highlighting the keywords “KGB chairman” users could easily locate the year as shown in the zoom of Figure 6. In this case, “KGB chairman” must be entered as query keywords for the highlighting to help in the skimming process.

As shown in this usage scenario, by reorganizing the index entries, the user can narrow down the number of entries that one must search through to find the correct answer. By entering all of the relevant keywords, users can see in one single screen what might be relevant without having to consult multiple index entries dispersed through several different index pages.

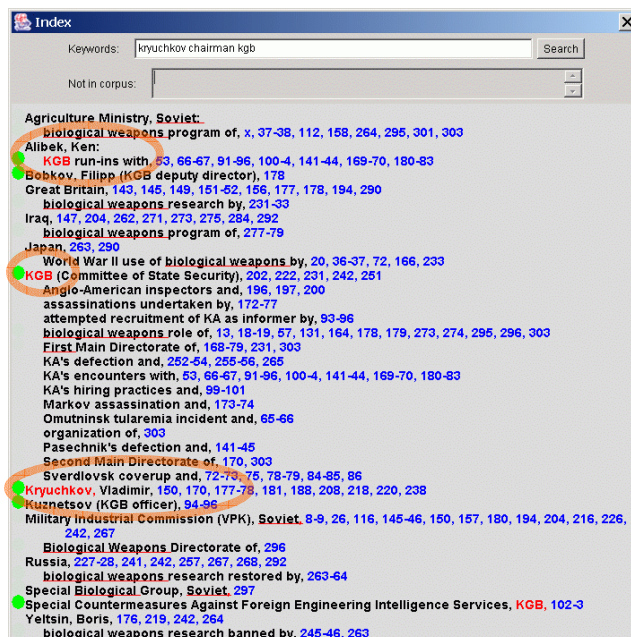


Figure 5: ScentIndex showing the reorganization after “kryuchkov chairman kgb” was entered.

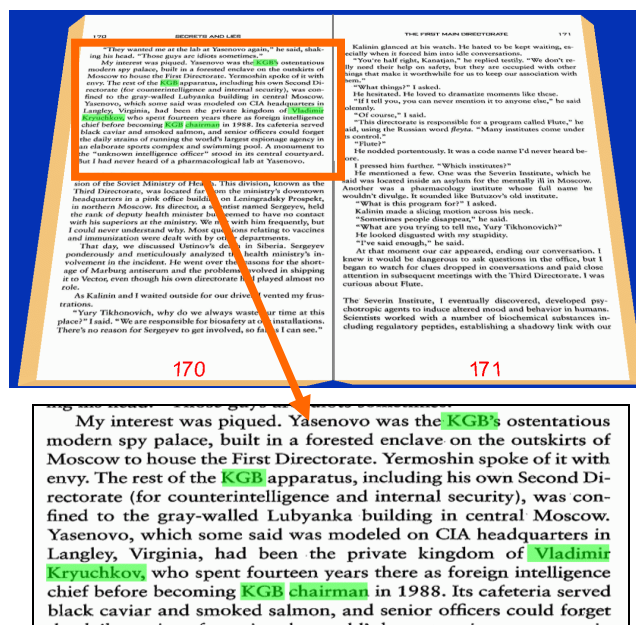


Figure 6: Clicking on the 2nd entry of “Kryuchkov, Vladimir” page 170, the book opens up to that page (top) and highlights any words in the list of “kryuchkov chairman kgb Vladimir”. The automatic highlighting helped in locating the relevant passages. Note that we automatically highlight not just the search keywords specified by the user, but also the words that were in the index entry, such as “Vladimir”.

Summary of User Study

Currently the paper version is the predominant way for efficient access of information contained in books. We have conducted a user study comparing the performance of ScentIndex to the paper version of the subject index. In our

companion manuscript [7], we describe the detail of the user study and its results.

Before the study, we were not sure if the conceptual reorganization of the index would hinder or speedup users. Would they get confused by the conceptual reorganization? Would they find the interaction cumbersome? Therefore, we are interested in evaluating the entire ScentIndex system for *retrieving, comparing, and comprehending* information contained in the subject index of a book as compared with the same subject index in the paper version. We studied these three task types for the speed and accuracy performance of both content experts and novices.

We found that users were *faster* in finishing their tasks and more *accurate* in their answers using the ScentIndex, regardless of their expertise level in the book content. Users also overwhelmingly preferred the ScentIndex interface for these tasks. The analysis results show that the interface condition did not have any interactions with the expertise level.

CONCLUSION

Reading is essential to the improvement and prolonging of human collective knowledge. The subject index, invented in the 15th century, has been one of the most important techniques invented to improve the retrieval, comparison, and comprehension of conceptual ideas in books. In this short paper, we described the usage scenario of a new way of using subject indexes in electronic books. ScentIndex is a method that conceptually reorganizes large subject indexes according to some information need. By reorganizing and reducing the index entries down to a single page, users can more efficiently navigate and scan for information of interest. We integrate this technique with highlighting and navigational enhancements in the eBook browsing interface to enable quick scanning and skimming of relevant passages. By taking advantage of existing subject indexes, we hope to preserve the look and feel of subject indexes in their electronic form as well as enhance them for actual use.

ACKNOWLEDGMENTS

The user study portion of this research has been funded in part by contract #MDA904-03-C-0404 to Stuart K. Card and Peter Piroli from the Advanced Research and Development Activity, Novel Intelligence from Massive Data program. We thank Jock Mackinlay for some fruitful conversation about the interaction of the eBook.

REFERENCES

1. Adobe. What is Adobe PDF?
<http://www.adobe.com/products/acrobat/adobepdf.html>, Retrieved Dec. 2003.
2. Alibek, Ken, Handelman, Stephen. *Biohazard*. Delta Publishing, New York, NY, 1999.
3. Amazon.com. Search Inside the Book.
<http://www.amazon.com/exec/obidos/tg/browse/-/10197021/ref%3Dsib%5Fmerch%5Fgw/104-3136902-3410324>. Retrieved December, 2003.
4. Bush, V. *As we may think*. The Atlantic Monthly 176, 1 (July 1945), 101--108.
5. Card, S. K., Robertson, G. G., & York, W. The WebBook and the Web Forager: An Information Workspace for the World Wide Web. In *Proc. of Human Factors in Computing Systems (CHI 96)*, pp. 111-117. ACM Press, 1996.
6. Card, S. K., Hong, Lichan, Mackinlay, Jock D., Chi, Ed H. 3Book: A 3D Electronic Smart Book. In *Proc. of the Advanced Visual Interfaces (AVI)*. (to appear), 2004.
7. Chi, Ed H., Lichan Hong, Julie Heiser, Stuart K. Card. ScentIndex: Conceptually Reorganizing Subject Indexes for eBooks. (submitted for publication). 2004.
8. Chi, E.H., Piroli, P., Chen, K., and Pitkow, J. Using information scent to model user information needs and actions on the Web. *Proc. of the ACM Conference on Human Factors in Computing Systems, CHI 2001* (pp. 490-497). ACM Press, 2001. Seattle.
9. CMU. The Million Book Project.
http://www.library.cmu.edu/Libraries/MBP_FAQ.html, 2003.
10. Cutting, Douglass, David Karger, Jan Pedersen, and John W. Tukey. (1992) Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proc. of the 15th Annual International ACM/SIGIR Conference*, Copenhagen.
11. Harrison, B. L. E-books and the future of reading. *IEEE Computer Graphics and Applications*, 20(3):32--39, May/June 2000.
12. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. Lexically-generated subject hierarchies for browsing large collections. *Int. Journal on Digital Libraries*, Vol. 2, No. 2/3, pp. 111-123. 1999.
13. Remde, J.R., Gomez, L.M., and Landauer, T.K. SuperBook: An automatic tool for information exploration - hypertext? In *Proc of Hypertext '87*, pp. 175-188. ACM Press, 1987.
14. Rocket eBook. www.rocket-ebook.com, 2003.
15. Schuetze, H., Manning, C. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press, 1999.
16. Silicon Graphics, "Demo Book," Silicon Graphics, Mountain View, California, Computer program 1993.
17. SoftBook. www.softbook.com, 2003.